EARLIER SCORES COUNT MORE: CAUSAL ANCHORING IN MULTI-ATTRIBUTE LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-as-a-judge protocols often score multiple criteria in a single response and aggregate them using an unweighted mean. Criterion order is known to shift scores, but correlation between earlier and later scores does not distinguish anchoring from shared dependence on item quality. We test for self-anchoring—the influence of a judge's own earlier score tokens on its later scores—by intervening directly on generated score tokens. Holding the item, rubric, criterion order, and other preceding scores fixed, we set an earlier score above or below its baseline and measure the expected later score. Across five open-weight judges evaluated on teledermatology images and text, the mean controlled direct effect is positive for every judge and dataset. In both text settings, knocking out score-to-score attention removes the positive effect, while blocking intermediate positions does not. This dependence is concentrated in late decoder layers. Replacing earlier scores with a fixed reference value removes 97–98% of variance in aggregate scoring across criterion permutations. Measuring each position's total effect on the aggregate, with later scores generated sequentially, shows that the first output position carries roughly $5\times$ its nominal weight in the smaller judges. Scoring one criterion per call improves agreement with dermatologists for most image criteria, but not for the aggregates, and lowers correlation with human helpfulness on text. Equal nominal weights therefore do not imply equal criterion influence, and removing self-anchoring may not necessarily improve the validity of aggregate scoring.

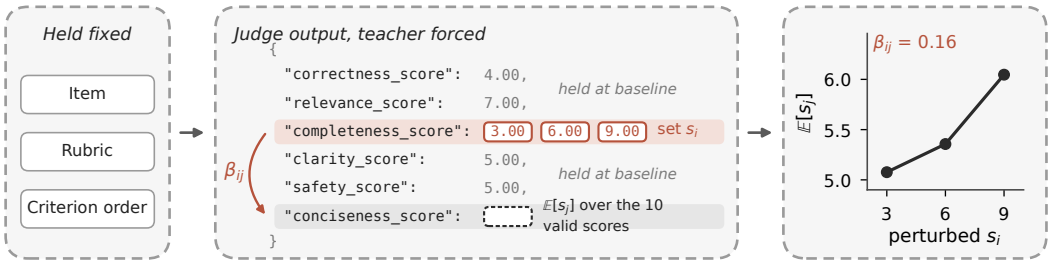
1 INTRODUCTION

LLM-as-a-judge evaluation uses language models to evaluate the outputs of other systems (Zheng et al., 2023). In rubric-based evaluation, a judge often produces one score for each of K criteria in a single output, and the scores are averaged into an aggregate score that is used to select, filter, and reward model outputs as though each criterion counted equally. Because the judge generates autoregressively, every score it generates is conditioned on the scores before it.

Stureborg et al. (2024) found that when GPT-4 scored the coherence and consistency of SummEval summaries (Fabbri et al., 2021) in one output, its two scores correlated at $r = 0.979$, compared with $r = 0.315$ for human experts. They also found that GPT-3.5-Turbo agreed less with experts on coherence when coherence was scored later, and they attributed this anchoring effect to autoregressive generation. Feuer et al. (2025) showed that LLM judges implicitly reweight the criteria they are given, and Xu et al. (2026a) showed that the order of criteria within a rubric shifts the scores. Prior interventions establish sensitivity to evaluation format or response content, but they do not isolate the contribution of an earlier emitted score while holding the evaluated item fixed, so a later score may still follow an earlier one because both reflect the quality of the item. The open questions are how large this dependence is, where in the model it arises, and how much weight it gives each criterion.

In this paper, we study *self-anchoring*: the causal influence of a judge's own earlier score tokens on its later scores, and its contribution to the aggregate score. We keep the item, rubric, and criterion order fixed, set the score for criterion i to a baseline value and to three points above and below it, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Intervention on one generation (b) Direct-effect estimate

Figure 1: The intervention. (a) With the item, rubric, and criterion order fixed, every score before the target is held at its baseline, the score at position i is set to its baseline and to three points above and below it, and the expected score at a later position j is computed from the token probabilities over the ten valid scores. (b) The controlled direct effect β_{ij} is estimated by the slope of the expected later score on the perturbed score, shown for the effect of completeness on conciseness in one Setting 1 response, whose baselines and perturbed values appear in (a).

compute the model's expected score for a later criterion j from its token probabilities (Figure 1). We run these experiments with five open-weight judges with 1.5B to 27B parameters on teledermatology photographs and on two text datasets.

Our contributions are:

- **Controlled interventions on score tokens.** We isolate controlled direct effects of earlier scores while holding the evaluated item and other preceding scores fixed. Across judges and datasets, 71–99% of controlled-direct-effect estimates are positive (Section 4.2).
- **A score-to-score attention pathway.** Attention knockout identifies dependence on earlier value tokens, concentrated in late decoder layers; fixing earlier scores removes 97–98% of aggregate-score variance across orders in both text settings (Sections 4.3 and 4.4).
- **Effective weights and decision sensitivity.** We measure total effects on the aggregate and compare them with a linear-SEM reconstruction. Excess weight follows output position under order reversal. On images, median aggregate-score ranges across orders are 9–12 points for the larger judges and 25 for the primary 2B judge on a 10–100 scale; every primary-judge decision flips at the tested cutoffs (Sections 4.1 and 4.4).
- **Validation against human labels.** Isolated scoring improves several criterion scores without improving aggregate agreement, distinguishing equal criterion influence from predictive validity (Section 4.5).

2 Related Work

Position and order effects. LLMs use information at the beginning and end of long inputs more effectively than information in the middle (Liu et al., 2024). Pairwise judges show position bias, so swapping the two candidates can change the verdict (Zheng et al., 2023), and averaging over both orders is the usual remedy (Wang et al., 2024a). Bias catalogues and benchmarks (Ye et al., 2025; Lai et al., 2026) measure evaluation biases by perturbing evaluation inputs. Norman et al. (2026) show that production judges can give consistent scores across repeated runs while remaining strongly biased by position. Closest to our work, Xu et al. (2026a) show that rubric judges favor score options at particular positions in a list and that criterion order within a rubric is a second source of bias. In their study, criterion position affected 58 of 60 (judge, criterion) tests, and score-option order alone changed the top candidate on 15–41% of prompts. Li et al. (2026) similarly found that reordering the levels within a rubric shifts scores. Each of these studies varies the prompt and measures how the scores change, which leaves open whether a later score changes because of its place in the rubric or because of the scores generated before it. The closest interventional design is that of Singh et al. (2026), who generate responses that satisfy or violate a target rubric criterion and measure how other, independently scored criteria respond. We instead hold the response fixed and

perturb the judge's own earlier score within a single output, which also lets us convert the controlled direct effects into an effective weight for each position.

Anchoring, entanglement and weighting. Stureborg et al. (2024) described the inflated correlation between attributes and the loss of agreement at later positions as anchoring. External anchoring also operates across calls: Kapetanovic et al. (2026) showed that a prior score supplied in the prompt shifts a judge's scores toward it. Our self-anchoring intervention instead targets a score token within the judge's own output sequence. Other studies examine how the rubric itself couples criteria. Li et al. (2025) identified a "criteria entanglement bias" in which judges assign similar scores across dimensions, and Yu et al. (2026) report rubric interference, in which a criterion's verdict shifts with the other criteria in the rubric. In a related finding, Bagaria et al. (2026) show that scores can be partly predicted from the rubric text alone. These studies manipulate the prompt, the rubric, or the response, and none isolates the effect of a score the judge has already emitted, which is the pathway we test. For criterion weighting, Feuer et al. (2025) found that overall scores across evaluated models were almost perfectly correlated with style scores ($r \approx 0.999$) and concluded that judges "implicitly reweight" their criteria.

Interventional methods. Our design follows causal tracing (Meng et al., 2022), causal mediation analysis (Vig et al., 2020), and attention knockout (Geva et al., 2023), which blocks a later position from attending to earlier ones. These methods have recently been applied to LLM judges. Vasava & Jiang (2026) use causal tracing, the logit lens and head knockout on two judges fine-tuned for summarization and find that the rating emerges in the upper layers, and Xu et al. (2026b) locate seven judge biases in a few directions of the model's activations and steer them. Both studies examine how a judge processes the item it scores. We apply the same methods to a different part of the context, the scores the judge has already generated.

3 Experimental setup

3.1 Datasets and rubrics

An item is one photograph or one text response. Each item is scored on $K = 6$ criteria on a 1–10 scale. The aggregate score is ten times the mean of the six scores, so each criterion has a nominal weight of $10/K = 1.667$ points of the aggregate score per point of its own score.

Images. We use photographs from SCIN, the Skin Condition Image Network (Ward et al., 2024), a public, de-identified dataset of 10,408 images of skin conditions contributed with informed consent by adults in the United States. The rubric rates six aspects of image quality that matter in teledermatology, namely focus, lighting, framing, color fidelity, artifacts, and visibility (Appendix E). The main image experiments use two disjoint sets. Four judges score 20 photographs under 24 criterion orders, and all four score 96 other photographs for the intervention, which three of them also score under the same 24 orders. The agreement analysis uses a fixed subset of 480 cases with dermatologist gradability labels that contains the 96, and Table 3 lists every photograph set with its use.

Text, Setting 1. Setting 1 contains 30 short responses that we wrote for 10 user requests, one each for factual recall, arithmetic, coding, a technical question, an explanation, summarization, household advice, personal advice, business writing, and workplace writing. Each request has a strong response and two responses with a deliberate flaw in one named criterion (Table 4), and the rubric scores correctness, relevance, completeness, clarity, safety, and conciseness.

Text, Setting 2. Setting 2 contains 40 responses sampled from HelpSteer (Wang et al., 2024b), a public dataset of 37,000 assistant responses with human ratings of helpfulness, correctness, coherence, complexity, and verbosity. We score them with the Setting 1 rubric, so Setting 2 replicates the text results with a different judge on data we did not write.

3.2 Models

Four vision-language models score the photographs, Qwen2-VL-2B-Instruct (the primary judge, the only image judge small enough for every experiment to run on one consumer GPU), Gemma 4 12B, Ministral 3 14B, and Qwen3.8-27B. In the text settings, Qwen2-VL-2B-Instruct scores Setting 1

without image input, Qwen2.5-1.5B-Instruct scores Setting 2, and Gemma 4 12B also scores Setting 2 to test whether the effective weights hold in a larger judge from another model family. Appendix B lists checkpoints and hardware.

3.3 Prompts and scores

Each judge receives the rubric prompt in Appendix E and returns a JSON object with one key per criterion, and we change the criterion order by reordering the rubric sections and the JSON keys together. Both prompts also request an `overall_score` key after the six criteria, which we leave out of every analysis, so the aggregate score is computed from the six criterion scores alone. Following G-Eval (Liu et al., 2023), we use the expected score over the ten valid values with the JSON up to that key supplied by teacher forcing (Appendix E).

3.4 Perturbing an earlier score

To test whether an earlier score influences a later one, we fix the item, rubric, and criterion order and change only the score at an earlier position i (Figure 1a). Every other score before the later position j is held at its baseline: the valid score nearest the judge's expected score at that position when all positions before it are set to 5. We set score i to its baseline and to three points above and below it, clipped to the 1–10 scale, and read the judge's expected score at j from its token probabilities. We define the *controlled direct effect* β_{ij} as the least-squares slope, with an intercept, of this expected score on the actual inserted value of score i (Figure 1b). Holding the item and all other preceding scores fixed isolates the direct influence of score i on score j, with no path through intermediate scores. We estimate all 15 criterion pairs with $i < j$ on every item and report the mean over pairs and items as the mean controlled direct effect β. Appendix A compares perturbations of one and two points separately with three points in the seven-level reruns.

3.5 Order sensitivity

To measure how much criterion order changes the aggregate score, we score the same items under multiple orders and decompose the total variance into item, order, and item-by-order interaction components. A judge's *order sensitivity* is the share explained by the order main effect, excluding the interaction. For images we use 24 counterbalanced orders, weighted equally, in which each criterion appears in each position four times. Order sensitivity is distinct from β: a judge can anchor strongly yet show low order sensitivity if its scores vary widely across items.

3.6 Effective weights

The *effective weight* w_i of output position i is the change in the aggregate score per point of change in score i, including changes this causes in later scores. We average the controlled direct effects over items to form a $K \times K$ matrix B, with $B_{ij} = \beta_{ij}$ for $i < j$ and zero otherwise. Under a linear model over output positions, path analysis (Bollen, 1987) gives the total-effect reconstruction

$$w_i = \frac{10}{K} \sum_{j=1}^{K} \left[(I - B)^{-1} \right]_{ij}, \tag{1}$$

where $K = 6$, I is the identity matrix, and $[\cdot]_{ij}$ denotes the row-i, column-j entry. The weight reduces to the nominal $10/K$ when $B = 0$. This reconstruction assumes that effects combine linearly.

We also measure effective weights directly. We perturb each position's score by three points in either direction, clipped to 1–10, while keeping earlier scores at baseline. We compute later score distributions sequentially and insert each modal score into the context. The weight is the difference between the two resulting aggregate scores, computed from expected later scores, divided by the actual difference between the inserted scores. We average these weights over items. We estimate B in each text setting's canonical order and measure weights in both canonical and reversed orders.

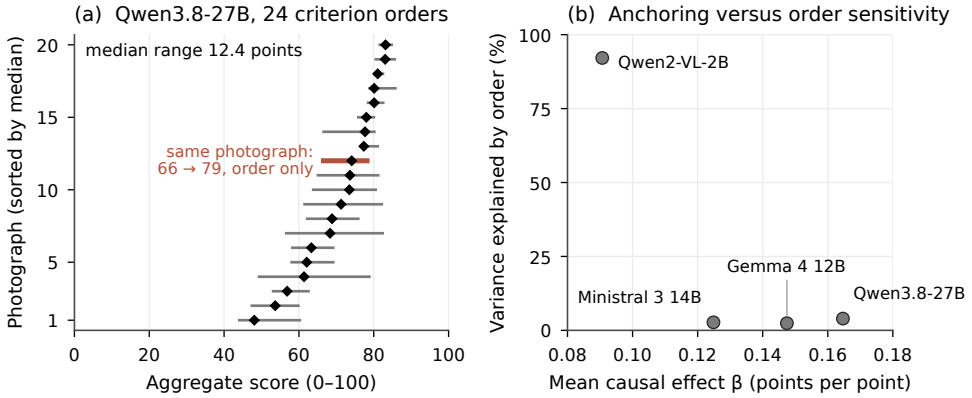


Figure 2: Criterion order changes the aggregate score. (a) Range (line) and median (diamond) of the aggregate score across 24 criterion orders for each of 20 photographs, Qwen3.8-27B. (b) Mean controlled direct effect β against the share of the variance of the aggregate score explained by order, one point per image judge.

4 Results and Analysis

4.1 Criterion order effects

To measure order sensitivity, four judges score the same 20 photographs under 24 permuted criterion orders in which each criterion appears in each position four times. Order changes the aggregate score of the same photograph for every judge (Figure 2). The primary judge's aggregate scores vary little across photographs, and order explains 92% of the variance of its aggregate score, with a median range across orders of 25 points out of 100. The three larger judges have smaller median ranges of 9 to 12 points. The standard deviation across photographs is three to five times the standard deviation across orders, so order explains only 2 to 4% of their variance (Table 9), even though they anchor more strongly than the primary judge (Section 4.2).

On the 96 intervention photographs, with a retake cutoff at each judge's median aggregate score, the accept or retake decision flips across orders for 51% of photographs with Gemma 4 12B and 38% with Ministral 3 14B, falling to 16 and 17% at a cutoff of 60 (Table 10). For the primary judge, every decision flips at each cutoff we tested. Its aggregate score has an AUROC against dermatologists' gradability labels ranging from 0.60 to 0.74 across choices of the first criterion, a range of 0.14; the corresponding ranges are 0.05 for Gemma 4 12B and 0.01 for Ministral 3 14B. Flip rates by skin type depend on the cutoff. For Gemma 4 12B, photographs of Fitzpatrick types V–VI flip more often than types I–II at the median cutoff (59% against 50%) and at 70 (50% against 32%) but less often at 60 (14% against 21%), and Ministral 3 14B shows no consistent ordering.

Takeaway. Criterion order changes decisions for fixed photographs, including every primary-judge decision at the tested cutoffs.

4.2 Earlier scores influence later scores

Figure 3 shows that perturbing an earlier score shifts the scores generated after it in the same direction, on average, for every judge and dataset. On the 96 photographs, raising an earlier score by one point raises a later score by 0.091 points for the primary judge and by 0.147, 0.125, and 0.165 points for Gemma 4 12B, Ministral 3 14B, and Qwen3.8-27B (Table 5). The larger judges are also more consistent, with 99% of their 1,440 estimates and all 15 criterion pairs positive. For the primary judge, 71% of estimates and 11 of 15 pairs are positive, and its four negative pairs all have artifacts as the later criterion. The primary judge's estimated effect also depends on the baseline score. Most baselines lie at 6 to 8, and its effect for a perturbation of one point is 0.157 across all (item, criterion-pair, order) combinations but 0.029 on those far enough from the top of the scale to allow three points in either direction (Appendix A). The text judges anchor most strongly, at 0.206 in Setting 1 and 0.202 in Setting 2, and their effects change little with the size of the perturbation.

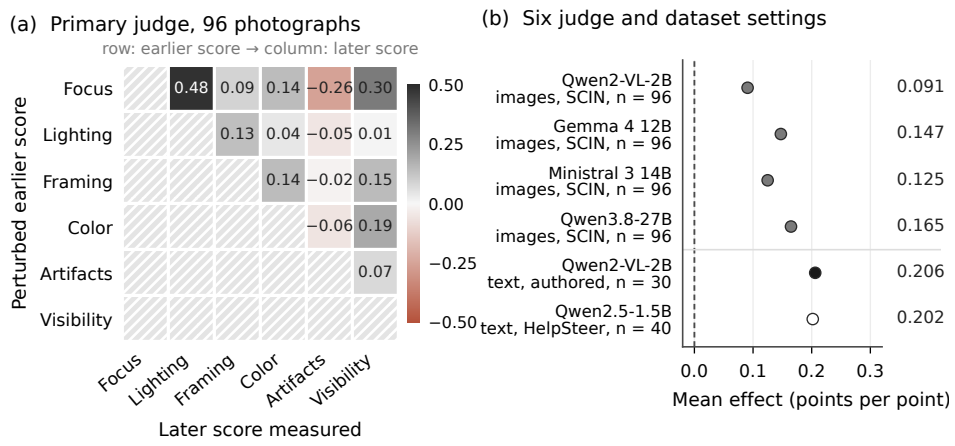


Figure 3: Earlier generated scores shift later scores. (a) Controlled direct effect β_{ij} averaged over items for each ordered criterion pair, primary judge, 96 photographs. (b) Mean controlled direct effect β over the 15 pairs for each judge and dataset, with bootstrap 95% intervals.

Takeaway. The mean controlled direct effect is positive for every judge and dataset, demonstrating self-anchoring beyond shared item quality.

4.3 Later scores attend to earlier score tokens

To test whether the effect depends on attention to the earlier score tokens, we apply attention knockout (Geva et al., 2023) in the canonical order of both text settings. The knockout adds a large negative value to the attention logits from chosen positions to the tokens of the perturbed score (Appendix E). We block attention from the positions that generate the later score, from the positions between the two scores, or from every later position within one third of the decoder layers. Blocking attention from the positions that generate the later score leaves a small negative residual, approximately -4% of the original effect in both settings (Figure 4a), whereas blocking the positions between the two scores leaves it intact (101% and 110%). Across layers, the effect depends mostly on the last third of the decoder, where knockout leaves 34% and 26% of it, compared with 80% and 76% for the middle third. The effect therefore depends on direct attention to the earlier score tokens, mostly in late layers. At the unperturbed baseline, the masks change the later score's expected value by 0.11 to 0.36 points in mean absolute value and its normalized entropy by at most 0.024 (Appendix E).

As controls, we replace the judge's earlier scores with those of another item, which keeps the form of the context while removing its relation to the item, and we set every earlier score to 5. Substitution changes the aggregate score of individual items by 2.9 and 3.7 points in mean absolute value, with no systematic direction, and setting every earlier score to 5 lowers the aggregate score by 2.76 and 5.17 points (Figure 4b, c).

Takeaway. Self-anchoring depends on score-to-score attention, with the strongest layer-third knockout effect in late layers.

4.4 Effective weights and primacy

Figure 5 shows that the first output position receives far more than its nominal weight of 1.667 in both text settings. In Setting 1, a change of one point in the first score changes the aggregate score by 8.25 points, and a change of one point in the second score changes it by 7.13. Setting 2, with a different judge and dataset, shows the same pattern with 8.04 and 4.63. Summed over positions, the weights reach 24.7 and 21.3 against a nominal 10. When we reverse the criterion order, the criterion moved from last to first receives a weight of 8.24 in Setting 1 and 7.47 in Setting 2, so the excess weight follows the output position whichever criterion occupies it.

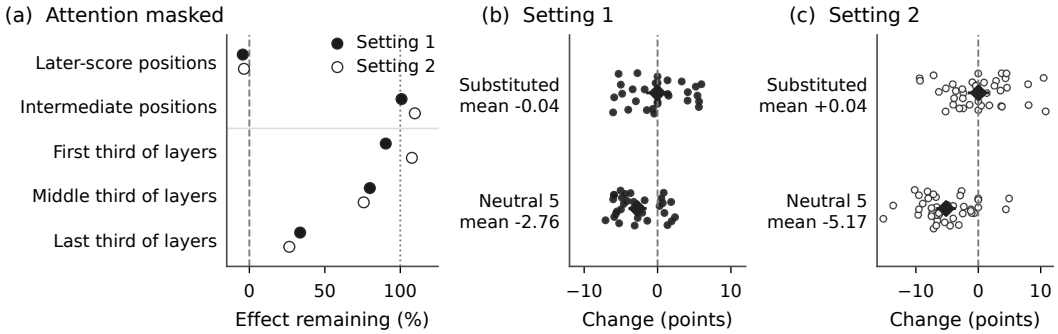


Figure 4: (a) Share of the effect of an earlier score on a later score that remains when attention to the perturbed score is blocked. (b, c) Change in the aggregate score relative to the own-baseline condition, in which every earlier score holds the response's own baseline, when another response's baselines are substituted and when every earlier score is set to 5. Points are responses, and diamonds are means with bootstrap 95% intervals.

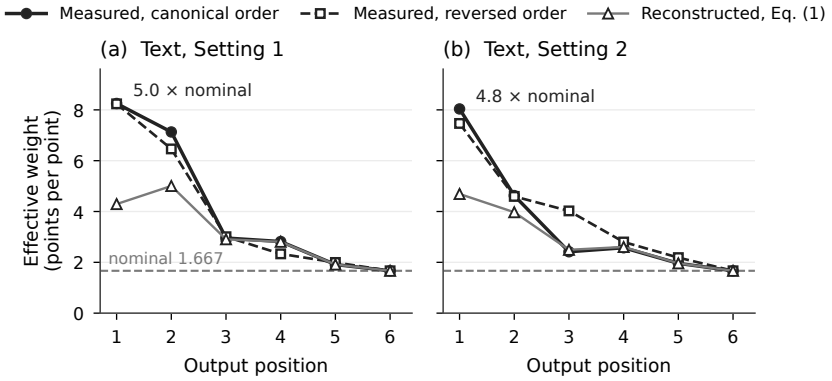


Figure 5: Effective weight by output position in (a) Setting 1 and (b) Setting 2. We measure weights by perturbing the score at a position by three points in either direction and computing subsequent score distributions sequentially, inserting each modal score into the context. The reconstruction applies Eq. (1) to the controlled direct effects in the canonical order. The dashed line is the nominal weight of 1.667.

In Gemma 4 12B, a larger judge from another model family scoring the same 40 HelpSteer responses, the excess is smaller. The first position receives a weight of 2.95 (95% CI 2.49 to 3.46) in the canonical order and 2.80 (2.25 to 3.62) in the reversed order, about 1.8 times the nominal weight. Summed over positions, the weights reach 12.9 and 15.1 (Table 8). In the reversed order, the first four positions all have weights of 2.8 to 2.9.

For the two smaller judges, Eq. (1) reproduces the measured weights at positions 3 to 6 but underestimates them at positions 1 and 2, giving 4.29 and 5.00 against the measured 8.25 and 7.13 in Setting 1 (Table 8). Most of this gap comes from propagating discrete modal scores rather than expected scores. A small change in a later score's expected value can shift its modal value by a whole point. The most likely value of a later criterion changes up to 2.6 times as steeply as its expected value, and accounting for this propagation recovers 7.04 of the measured 8.25 at position 1 and 6.03 of 7.13 at position 2. The remainder reflects interactions between scores. When the first two scores, correctness and relevance, are perturbed together in the same direction, the effect on completeness is 0.11 points per point smaller than the sum of their separate effects, and the effect on safety is 0.08 larger. The reconstruction misses both sources of propagation. The measured weights characterize sequential scoring with modal scores inserted into the context.

Across the permuted orders, the effect of an earlier score also weakens the later it appears in the output. We regress each estimated effect on the position of the earlier criterion and the distance

between the two criteria, controlling for the identity of both criteria and computing p-values with a wild cluster bootstrap (Cameron et al., 2008). The effect falls by 0.008 per position for images and by 0.050 and 0.037 per position in the two text settings ($p < 10^{-4}$, Table 6).

Effective weights also predict how sensitive each position is to criterion order, tracking the SD across orders at each position with $r = 0.95$ in Setting 1 and $r = 0.99$ in Setting 2. Replacing the earlier generated scores with a fixed score of 5 in each of the seven orders we ran reduces the variance across orders by 97.5% in Setting 1 and 98.2% in Setting 2, so nearly all of the order sensitivity in these settings depends on the earlier score tokens.

Takeaway. Equal nominal weights conceal unequal total effects, and excess influence follows output position under order reversal.

4.5 Agreement with human labels

We compare joint, isolated, and corrected scoring against human labels on two datasets. On HelpSteer, Qwen2.5-1.5B-Instruct scores 200 responses, including the 40 of Setting 2, in three conditions. The joint condition scores all six criteria in one generation, and the isolated condition scores each criterion in first position of the same rubric with no earlier scores. The corrected condition subtracts from each joint score the anchoring predicted by the controlled direct effects of Setting 2. For criterion j this is $\sum_i \beta_{ij}(s_i - \bar{s}_i)$ over the earlier criteria i, where s_i is the joint expected score and $\bar{s}_i$ is its mean over responses. We report the Spearman correlation of each aggregate score with human helpfulness and of each criterion with its matching human attribute. On SCIN, the primary judge scores all 480 photographs of our subset, including the 96 used for the intervention, both jointly and with each criterion alone in its own call. We correct the joint scores with the controlled direct effects of the 96 photographs. We report the AUROC of each score against the majority gradability label of three dermatologists and test differences with a paired bootstrap over responses or photographs.

On HelpSteer (Table 11), the joint aggregate score correlates with helpfulness at $\rho = 0.254$, and scoring each criterion in isolation significantly lowers this to 0.113 (difference -0.140, 95% CI -0.234 to -0.051). The corrected score (0.218) does not differ significantly from the joint one. Clarity, scored fourth, agrees with human coherence at 0.218 jointly but only 0.088 in isolation. Conciseness, scored last, agrees with human verbosity at 0.112 jointly, 0.163 in isolation, and 0.321 after correction. Without the 40 responses of Setting 2, whose controlled direct effects define the correction, the aggregate score gives 0.313 jointly, 0.170 in isolation, and 0.279 after correction.

On SCIN the joint aggregate score reaches an AUROC of 0.707. One criterion per call gives 0.727 at 2.3 times the input tokens (difference $+0.020$, 95% CI -0.026 to $+0.066$), and correction gives 0.687 (difference -0.019, 95% CI -0.035 to -0.005). Scoring one criterion per call raises the AUROC of five of the six criteria, most for artifacts (0.560 to 0.700), and lowers it only for color (0.687 to 0.607). Correction lowers lighting to 0.488 because it subtracts the expected focus score, although the judge generated the same focus value for 461 of the 480 photographs. Applying the correction to the values that enter the context instead leaves lighting at 0.678 and the aggregate score at 0.707 (Table 11). For the three larger judges, the AUROCs of joint and isolated aggregate scores differ by 0.01 or less on 60 photographs, a subset too small to rule out meaningful differences.

Takeaway. Improving criterion-level agreement by isolated scoring does not guarantee better aggregate agreement.

5 Discussion

Our results show that judges attend to earlier score tokens and give early positions more than their nominal weight. The concentration of the effect in the last third of the decoder layers also fits the finding of Vasava & Jiang (2026) that judges fine-tuned for evaluation form their rating in the upper layers.

This mechanism is consistent with the anchoring effect reported by Stureborg et al. (2024), since dependence on earlier score tokens would produce both the inflated correlation between attributes scored in one output and the lower agreement with experts for attributes predicted later. It is also consistent with the criterion order effects of Xu et al. (2026a), since holding earlier scores at a

fixed reference value removes 97–98% of the variance across orders in our text settings. The rubric interference of Yu et al. (2026) may arise the same way when criteria are scored in sequence.

Our comparison with human annotations qualifies the recommendation of Stureborg et al. to score one attribute per generation, which they based on evidence that anchoring reduces agreement with experts. In our experiments, scoring one criterion per call improved agreement for most individual criteria, and we detect no improvement for the aggregate score. As a measurement of six nominally distinct criteria, the joint aggregate score is biased, because its effective weights differ from the intended weights and depend on criterion order. As a predictor of human judgments, however, it outperforms isolated scoring on HelpSteer, and on SCIN we detect no difference between them, likely because earlier scores contain information about overall quality that is unavailable when each criterion is scored on its own. Joint scoring departs from the intended equal weighting but yields higher correlation with human helpfulness on HelpSteer. Stureborg et al. also noted that such dependence may be "natural (and in part valid)". Removing earlier scores reduces correlation with human helpfulness by 0.140.

For deployed judges, scoring one criterion per call is preferable when the criterion scores themselves are the output of interest. An image-quality screening system that provides criterion-specific feedback, for example, depends on exactly the criterion scores that isolated scoring improved most in our image experiments. Permuting the criterion order and averaging can remove the positional weighting from an aggregate computed over many items, but it does not stabilize a single judgment, such as the decision to accept or reject one photograph. In a teledermatology intake workflow, the criterion placed first in the schema can shape the effective decision rule even when all criteria have equal nominal weight. If that positional effect interacts with skin tone, the rubric alone will not reveal it. In our data the skin type pattern depends on the cutoff for Gemma 4 12B and is absent for Ministral 3 14B (Section 4.1), and with 22 to 34 photographs per group we cannot separate it from noise.

We therefore recommend that practitioners who validate a deployed judge report the criterion order, the effective weights, and the flip rate across plausible orders at the cutoff they use, including for relevant subgroups. Effective weights can be estimated with the procedure of Section 3.6 and are most useful as a diagnostic of how the judge actually combines its criteria. In our experiments, correction improved agreement for the last text criterion with no detectable improvement in the aggregate score (Section 4.5), so the weights are better suited to diagnosis and rubric design than to correction. When one criterion should drive the decision, it can be placed first. When the goal is to limit how far scoring errors spread, the criteria the judge scores least reliably can be placed last, where they influence the fewest later scores.

Finally, among the image judges, mean controlled direct effect and order sensitivity diverge with size. The 12B, 14B, and 27B judges anchor more strongly than the 2B judge, although part of the 2B judge's weaker effect may come from its baseline scores (Section 4.2), while the photograph explains 87–92% of the variance of their aggregate scores, compared with 3% for the 2B judge. A judge whose scores vary more across items can therefore anchor strongly while order accounts for little of the variance in its aggregate score. Order still changes an individual photograph's score by about 10 points, which matters most near a cutoff, where small score changes can alter the decision.

6 Conclusion

We studied how the scores an LLM judge generates within a single multi-attribute judgment depend on one another. By perturbing earlier score tokens, we find that later scores depend on earlier ones through attention to their tokens and that early output positions receive more than their nominal weight, so equal nominal weights do not produce an equally weighted aggregate score. Among the image judges, larger judges anchor more strongly, and order accounts for a smaller share of the variance in their aggregate scores because their scores vary more across photographs, although it still changes many decisions near a decision cutoff. Because removing the dependence changes agreement with human annotations in both directions, the choice between joint scoring and one criterion per call should depend on whether the intended output is a set of criterion scores or an aggregate score. Criterion order is an evaluation design choice that should be reported with results, together with effective weights, whenever an aggregate score is used to select, filter, or reward outputs.

LIMITATIONS

Open models only. The intervention needs token probabilities and attention masking, so it cannot be run on the proprietary judges most used in practice.

Judges and settings. The aggregate scores of the 2B primary judge vary little across photographs, so the range of 25 points mostly reflects criterion order. The controlled direct effects on text and knockout results come from 1.5B and 2B judges, the effective weights from these judges and Gemma 4 12B on Setting 2, and the other larger judges were run on images only. The knockout localizes the effect to a third of the layers without identifying individual heads; head-level attribution remains future work. Agreement with human annotations rests on one text judge and one image judge, and the skin type groups contain 22 to 34 photographs each.

Size of the perturbation. The perturbed values sit three points from the baseline, which may place the context outside the range of scores the judge usually generates. With perturbations of one and two points the mean effect stayed positive (Appendix A), but the slope depended on the size of the perturbation and on the items included. For Gemma 4 12B a perturbed value reached the end of the scale in 182 and 141 of 240 (item, source-position) interventions, so its weights reflect smaller realized changes.

Controls and variants. We did not mask an equal number of other tokens, so part of the knockout result could reflect disruption from masking itself, although the masks barely changed the unperturbed scores. Blank or shuffled images, which would show how much each image judge relies on the photograph, were not scored, and the rubric sections and JSON keys always moved together, so rubric position and output position are not separated. All scores are expected values computed from the model's score probabilities, and we did not compare sampled scoring, reasoning before scoring, or both text judges on both text datasets. The overall score both prompts request is not analyzed, and the fixed reference score of 5 used as a control can itself act as an anchor.

ETHICS STATEMENT

The image data are public, de-identified photographs from SCIN (Ward et al., 2024), contributed through a digital informed consent process and used under the SCIN Data Use License. We used no private clinical data, patient images, or personally identifying information, and made no attempt to identify any contributor. The text data are responses written by the authors and public responses from HelpSteer, and the study involved no interaction with human participants.

REPRODUCIBILITY STATEMENT

The perturbation, the effective weights, and the attention knockout are defined in Sections 3.4, 3.6, and 4.3, and the sensitivity to perturbation size is reported in Appendix A. Appendix B lists every judge with its pinned checkpoint revision, software versions, precision, image preprocessing, chat template, and hardware. Appendix C describes the datasets with an example, and Appendix D gives the numbers behind the figures. Appendix E reproduces the exact rubric prompts, with the SHA-256 digest of the image prompt, and describes how scores are computed from token probabilities and how the knockout mask is applied. The run records and analysis scripts are available from the authors on request.

AI USE STATEMENT

In this work, Claude (Anthropic) was used under the authors' direction to review code used for analyses and figures. Codex (OpenAI) was used to format and edit prose. No AI tool ran an experiment and every number in the paper was validated by the authors. We have reviewed all work produced with these tools and take responsibility for the final content, including text, claims, and artifacts.

REFERENCES

Anshul Bagaria, Sowmya S. Sundaram, Gokul S. Krishnan, and Balaraman Ravindran. Judging LLM-as-a-judge: Concerning rubric artifacts in LLM-based automated text generation evaluation.

In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. arXiv:2609.02942.

Kenneth A. Bollen. Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, 17:37–69, 1987. doi: 10.2307/271028.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008. doi: 10.1162/rest.90.3.414.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P. Dickerson. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, 2023. Association for Computational Linguistics.

Ante Kapetanovic, Kemal Altwlkany, Andro Mercep, Tomislav Duricic, and Emanuel Lacic. Anchoring bias in LLM-as-a-judge systems: Prior scores compromise evaluation independence. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*. Association for Computing Machinery, 2026. doi: 10.1145/3799682.3840879.

Peng Lai, Zhihao Ou, Yong Wang, Longyue Wang, Jian Yang, Yun Chen, and Guanhua Chen. BiasScope: Towards automated detection of bias in LLM-as-a-judge evaluation. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. Evaluating scoring bias in LLM-as-a-judge. In *Database Systems for Advanced Applications (DASFAA 2026)*, 2026. arXiv:2506.22316.

Weiyuan Li, Xintao Wang, Siyu Yuan, Rui Xu, Jiangjie Chen, Qingqing Dong, Yanghua Xiao, and Deqing Yang. Curse of knowledge: When complex evaluation context benefits yet biases LLM judges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2509.03419.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias, 2026. arXiv:2606.19544.

Divyansh Singh, Reza Davari, and Afra Mashhadi. RADAR: Rubric-aware dependency and redundancy analysis for LLM-as-judge evaluation, 2026. arXiv:2608.01810.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. arXiv:2405.01724.

Himil Vasava and Ming Jiang. Beyond scores: Understanding LLM-as-a-judge mechanisms in summarization evaluation, 2026. arXiv:2609.01604.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3384, Mexico City, Mexico, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.185.

Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Margaret Ann Smith, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, Christopher Semturs, and Pooja Rao. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7 (11):e2446615, 2024. doi: 10.1001/jamanetworkopen.2024.46615.

Yuzheng Xu, Tosho Hirasawa, Tadashi Kozuno, and Yoshitaka Ushiku. Am I more pointwise or pairwise? Revealing position bias in rubric-based LLM-as-a-judge, 2026a. arXiv:2602.02219v3.

Zixiang Xu, Sixian Li, Huaxing Liu, Xiang Wang, Shuai Li, Zirui Song, and Xiuying Chen. Inside the unfair judge: A mechanistic interpretability account of LLM-as-judge bias, 2026b. arXiv:2607.11871.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? Quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Dingyao Yu, Tong Zhang, Yutao Mou, Yunxiao Zhang, Wei Ye, and Shikun Zhang. Mitigating rubric interference in LLM judges via on-policy self-distillation, 2026. arXiv:2608.14684.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

A Sensitivity to perturbation size

To check whether anchoring persists under smaller perturbations, we reran the primary image judge and both text settings on the same items, setting the earlier score at seven levels from three below to three above its baseline. For each (item, criterion-pair, order) combination we compute the symmetric slope at radius h,

$$\frac{\mathbb{E}[s_j \mid s_i = b_i + h] - \mathbb{E}[s_j \mid s_i = b_i - h]}{2h},$$

where b_i is the baseline of earlier score s_i, s_j is the later score, and $h \in \{1, 2, 3\}$ is measured in score points. The context stays fixed as in Section 3.4. We retain combinations allowing $h = 3$ on both sides within 1–10 (Table 1).

Table 1: Mean symmetric slopes on (item, criterion-pair, order) combinations allowing all three radii. Reruns use the same 96 photographs and 30 and 40 text responses as the original experiments, with one canonical image order and six cyclic text orders. Each combination is evaluated at all seven levels; the full sets contain 1,440, 2,700, and 3,600 combinations, respectively, with retained counts below. The 95% intervals use 20,000 percentile cluster-bootstrap resamples, clustered by photograph for images and user request for text.

Setting	Records	±1	±2	±3
Images, Qwen2-VL-2B	965	0.029 [0.025, 0.034]	0.062 [0.059, 0.065]	0.059 [0.057, 0.062]
Text, Setting 1	2,540	0.235 [0.229, 0.242]	0.228 [0.221, 0.234]	0.209 [0.203, 0.216]
Text, Setting 2	2,992	0.207 [0.199, 0.214]	0.203 [0.197, 0.209]	0.197 [0.191, 0.203]

The mean effect is positive at every radius in all three settings. In text, the ±1 slopes (0.235 and 0.207) exceed the ±3 slopes (0.209 and 0.197), so ±3 gives a conservative estimate relative to ±1 within these reruns. For the primary image judge, most baselines lie at 6 to 8 (Figure 6), and the ±1 slope is 0.157 across all combinations but 0.029 among combinations allowing ±3. This dependence on baseline position is consistent with ceiling compression in the primary judge.

These runs used a different CUDA build (12.4 versus 12.1 for the original text runs), with some rounded baselines changed, so we compare radii only within the extended runs and do not pool them with the main estimates.

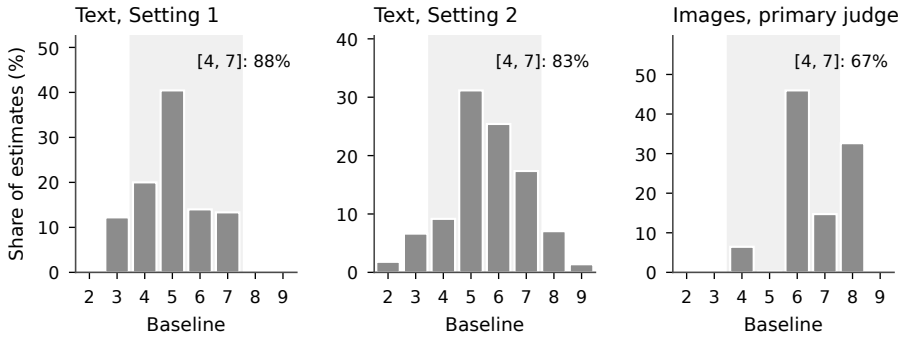


Figure 6: Distribution of the baseline of the perturbed score. The shaded band marks baselines for which both ±3 perturbations lie within the 1–10 scale.

B MODELS, CHECKPOINTS AND COMPUTE

Table 2 lists every judge with the checkpoint revision the runs were pinned to. The primary judge and the two small text judges ran on an NVIDIA RTX 3070 8 GB GPU with PyTorch 2.5.1 and Transformers 4.57.1, and the larger judges ran on NVIDIA H100 80 GB GPUs with PyTorch 2.13 and Transformers 5.15. On 96 photographs, the pairwise intervention took about 50 minutes for Qwen2-VL-2B on the RTX 3070, about 19 minutes each for Gemma 4 12B and Ministral 3 14B, and about 2 hours for Qwen3.8-27B. Scoring 24 criterion orders took about 10 minutes for Gemma 4 12B and Ministral 3 14B and 68 minutes for Qwen3.8-27B. The text pairwise experiments took 17 to 40 minutes depending on hardware, and the Gemma 4 12B effective weights took about four minutes per order. The extended runs used the same checkpoints and the software of the small judges on H100 and A10G GPUs. The supplementary Qwen3-VL 4B and 8B checks in Table 9 used 4-bit quantized checkpoints served by Ollama with greedy decoding at temperature 0.

All judges ran in bfloat16, the dtype of each checkpoint, without quantization and with the default attention implementation of Transformers (SDPA), and the log softmax over candidates was computed in float32. Each photograph was cropped to a centred square and resized to 448 by 448 pixels, and for SCIN cases with several photographs we used the first. At this size Qwen2-VL-2B and Gemma 4 12B receive 256 image tokens, Ministral 3 14B receives 256 image tokens with 16 separator tokens, and Qwen3.8-27B receives 196. Prompts were built with each model's chat

template and a generation prompt, with thinking disabled, so the Gemma 4 and Qwen3.8 prefixes contain an empty thought block before the JSON. The Qwen templates add their default system turn, and the Ministral 3 template adds its long default system prompt. The judges also differ in attention layout, since Qwen3.8-27B (Transformers model type `qwen3_5`) has 48 Gated DeltaNet linear attention layers and 16 full attention layers. Gemma 4 12B alternates five sliding window layers (window of 1,024 tokens) with one global layer, so in its sliding layers the score tokens, which sit more than 1,300 tokens after the image, cannot attend to the image directly. Ministral 3 14B and the two Qwen2 models use full attention in every layer. The attention knockout of Section 4.3 applies directly only to full attention layers, which complicates extending it to the larger judges.

Table 2: Judges and the checkpoint revisions the runs were pinned to.

Judge	Revision	Params	Layers	Scored	Hardware
Qwen2-VL-2B-Instruct (primary)	`895c3a4`	2.21B	28	images and Setting 1 text	RTX 3070, H100, A10G
Qwen2.5-1.5B-Instruct	`989aa79`	1.54B	28	Setting 2 text and human agreement	RTX 3070, H100
Gemma 4 12B-it	`707f0a3`	11.96B	48	images and Setting 2 weights	H100 80 GB
Ministral 3 14B-Instruct-2512	`3cea74c`	13.95B	40	images	H100 80 GB
Qwen3.8-27B	`1d4bf0f`	27.36B	64	images	H100 80 GB

C Datasets and an example

Table 3 summarizes the datasets, including SCIN[1] and HelpSteer.[2] We wrote Setting 1 so that each request has one strong response and two responses with a named flaw, and Table 4 shows one request with its three responses.

Table 3: Datasets and their uses. The 96 intervention photographs are disjoint from the 20 used to study criterion order.

Set	Source	Items	Human label	Used in
Images, order	SCIN	20	–	24 orders, four image judges
Images, intervention	SCIN (Ward et al., 2024)	96	gradability (three dermatologists)	intervention, all image judges, and 24 orders for three judges
Images, agreement	SCIN, includes the 96	480	gradability	joint, isolated, and corrected scoring, primary judge
Images, agreement	SCIN	60	gradability	joint and isolated scoring, three larger judges
Images, scale check	SCIN	100	–	six orders, quantized Qwen3-VL 4B and 8B
Text, Setting 1	authored, 10 requests $\times$ 3	30	quality tier and flaw	intervention, knockout, weights
Text, Setting 2	HelpSteer (Wang et al., 2024b)	40	five human attributes	intervention, knockout, weights
Text, human agreement	HelpSteer, including Setting 2	200	human helpfulness	Section 4.5

Table 4: A Setting 1 request with its three responses. The mixed response has a completeness flaw and the weak response a correctness flaw.

Request	Which city is the capital of Australia, and why isn't it Sydney?
Strong	Canberra is the capital of Australia. When the colonies federated in 1901, Sydney and Melbourne both wanted the seat of government, and neither would accept the other. The compromise written into the constitution was a purpose-built capital in New South Wales but at least 100 miles from Sydney, with Melbourne serving as the temporary seat until it was ready. Canberra was chosen in 1908 and Parliament moved there in 1927.
Mixed	The capital of Australia is Canberra. Sydney is the largest city in the country and is often assumed to be the capital, but it is not.
Weak	Sydney is in fact the capital of Australia, and has been since federation, because it was the largest and oldest colonial settlement. Canberra is a smaller inland city that houses some government offices but has never held capital status.

D Additional results

Tables 5–9 give the numbers behind Figures 3–5 and Section 4.1.

[1] `https://github.com/google-research-datasets/scin`

[2] `https://huggingface.co/datasets/nvidia/HelpSteer`

Table 5: Controlled direct effect of an earlier score on a later score by judge, with bootstrap 95% intervals clustered by item.

Judge	Data	Items	Effects	β [95% CI]	Positive	Levels, centre
Qwen2-VL-2B	images	96	1,440	0.091 [0.089, 0.093]	71%	3, valid score
Gemma 4 12B	images	96	1,440	0.147 [0.142, 0.153]	99%	3, valid score
Ministral 3 14B	images	96	1,440	0.125 [0.121, 0.128]	99%	3, valid score
Qwen3.8-27B	images	96	1,440	0.165 [0.159, 0.171]	99%	3, valid score
Qwen2-VL-2B	text, Setting 1	30	2,700	0.206 [0.202, 0.210]	95%	3, valid score
Qwen2.5-1.5B	text, Setting 2	40	3,600	0.202 [0.197, 0.207]	93%	3, valid score

Table 6: Change in the effect of an earlier score per output position. The position coefficient $\hat{\gamma}$ controls for criterion identities and distance between criteria (Section 4.4); its 95% intervals and p-values use the wild cluster bootstrap of Cameron et al. (2008), suited to settings with few clusters ($p < 10^{-4}$ in both settings).

Measure	Setting 1	Setting 2
Effect per output position, $\hat{\gamma}$	-0.050 [-0.054, -0.047]	-0.037 [-0.041, -0.032]

Table 7: Share of the controlled direct effect remaining under each attention knockout, with bootstrap 95% intervals over responses. The layer thirds are layers 0–8, 9–17, and 18–27.

Positions blocked from the perturbed score	Setting 1 (%)	Setting 2 (%)
Only the positions generating the later score	-4.4 [-5.9, -2.8]	-3.6 [-6.9, -0.5]
Only the positions between the two scores	100.8 [99.4, 102.3]	109.6 [106.5, 113.0]
Every later position, first third of layers	90.4 [89.8, 91.0]	107.7 [105.9, 109.5]
Every later position, middle third of layers	79.8 [78.0, 81.6]	75.8 [73.8, 77.7]
Every later position, last third of layers	33.6 [31.6, 35.5]	26.4 [23.2, 29.6]

Table 8: Effective weight by output position (nominal 1.667), in the canonical order unless marked reversed. Gemma 4 12B scores the Setting 2 responses.

	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Pos. 6
Setting 1, measured	8.25	7.13	2.96	2.82	1.91	1.67
Setting 1, reconstructed	4.29	5.00	2.91	2.81	1.91	1.67
Setting 2, measured	8.04	4.63	2.42	2.57	1.96	1.67
Setting 2, reconstructed	4.69	3.97	2.49	2.60	1.96	1.67
Gemma 4 12B, measured	2.95	2.92	1.81	1.74	1.77	1.67
Gemma 4 12B, measured, reversed	2.80	2.78	2.88	2.84	2.12	1.67

Table 9: Aggregate score variation across criterion orders by judge, on 20 photographs unless stated. The quantized Qwen3-VL rows use six orders and greedy decoding, with variance shares on 100 photographs.

Judge	Orders	Median range	SD across orders	SD across photographs	Order share	Mean β
Qwen2-VL-2B	24	25.2	7.52	1.99	92%	0.091
Gemma 4 12B	24	10.3	3.03	10.23	2%	0.147
Ministral 3 14B	24	9.4	2.20	10.78	3%	0.125
Qwen3.8-27B	24	12.4	3.10	9.96	4%	0.165
Qwen2-VL-2B, 96 photographs	24	25.5	7.56	1.54	–	0.091
Gemma 4 12B, 96 photographs	24	9.9	2.74	11.73	–	0.147
Ministral 3 14B, 96 photographs	24	8.7	2.18	13.67	–	0.125
Qwen3-VL 4B (quantized)	6	5.0	1.85	5.09	1%	–
Qwen3-VL 8B (quantized)	6	5.0	2.48	5.84	3%	–

Table 10: Decision changes across 24 orders on the 96 intervention photographs, by cutoff and Fitzpatrick group (I–II $n = 34$, III–IV 25, V–VI 22). The last column is the range of AUROC against gradability across the six choices of first criterion.

Judge	Cutoff	All	I–II	III–IV	V–VI	AUROC by first criterion
Qwen2-VL-2B	median (72.7)	100%	100%	100%	100%	0.60–0.74
Qwen2-VL-2B	70 and 60	100%	100%	100%	100%	
Gemma 4 12B	median (75.2)	51%	50%	52%	59%	0.78–0.83
Gemma 4 12B	70	39%	32%	36%	50%	
Gemma 4 12B	60	16%	21%	4%	14%	
Ministral 3 14B	median (72.0)	38%	47%	28%	36%	0.81–0.82
Ministral 3 14B	70	30%	35%	16%	41%	
Ministral 3 14B	60	17%	18%	16%	18%	

Table 11: Agreement with human labels by scoring condition, as Spearman ρ with the matched attribute on 200 HelpSteer responses and AUROC against gradability on 480 SCIN photographs. The last column applies the correction to the modal scores inserted into the context, which are not stored for HelpSteer.

Score (position)	Human label	Joint	Isolated	Corrected	Corrected, modal
HelpSteer, Spearman ρ					
Correctness (1)	correctness	0.170	0.170	0.170	–
Clarity (4)	coherence	0.218	0.088	0.173	–
Conciseness (6)	verbosity	0.112	0.163	0.321	–
Aggregate	helpfulness	0.254	0.113	0.218	–
SCIN, AUROC against gradability					
Focus (1)	gradability	0.706	0.746	0.706	0.706
Lighting (2)	gradability	0.678	0.712	0.488	0.678
Framing (3)	gradability	0.715	0.744	0.687	0.715
Color (4)	gradability	0.687	0.607	0.650	0.691
Artifacts (5)	gradability	0.560	0.700	0.637	0.608
Visibility (6)	gradability	0.657	0.715	0.570	0.534
Aggregate	gradability	0.707	0.727	0.687	0.707

E PROMPTS AND SCORE PROBABILITIES

The exact rubric prompts are reproduced below. When criterion order was varied, we reordered both the rubric sections and the corresponding JSON keys so that each criterion and its output position moved together.

Reading a score. Scores were computed from the model's next-token probabilities under teacher forcing, and no scores were sampled. For a target criterion j, the assistant response was provided up to the target JSON key, leaving its value unspecified (Table 14). Earlier keys contained values written to two decimal places, such as `8.00`.

We evaluated ten candidate completions corresponding to the valid scores `1.00` through `10.00`. Each candidate was tokenized with the model's tokenizer, and we summed the log probabilities of the tokens corresponding to the numeric value. When tokenizer offset mappings were unavailable, we used the tokens added beyond the common prefix. The resulting sum is the log likelihood of the complete numeric value conditioned on the image, prompt, JSON prefix, and any preceding candidate tokens.

All tokenizers used in these experiments split the digits individually, so scores `1.00` through `9.00` span four tokens and `10.00` spans five. For the primary image judge, tokenizer offsets were unavailable, so the trailing comma, which follows all ten values, was included in every candidate likelihood.

We normalized the ten candidate log likelihoods with a softmax to obtain a distribution over valid scores, excluding probability assigned to other continuations. We report its mean as the expected score and its most probable value as the modal score. Entropy is normalized by its maximum over ten scores.

Although the prompt permits scores with two decimal places, this measurement evaluates only the ten integer scores written in decimal form, and probability assigned to intermediate values such as `7.50` is not included. This discretization reduces resolution but preserves a common set of score levels across models.

Implementation. The image tokens, prompt, and JSON prefix shared by all ten candidates were evaluated once using a key-value cache, and the ten candidate completions were then scored in a single batch against that cache. When an intervention modified positions within a candidate sequence, we instead scored each full sequence directly. For Qwen3.8-27B, rotary positions for the continuation were constructed from the cached prefix. For Qwen2-VL, multimodal rotary positions were continued using the stored position offsets.

Attention knockout. To test whether later scores depended on direct attention to an earlier score value, we masked attention from selected query positions to the tokens representing the perturbed score, such as `7.00`, by adding the minimum value of the model dtype to the attention logits before the softmax. The associated JSON key, colon, and all other score tokens remained visible. We applied the mask at the target score positions, which include the open slot and every candidate token, at the positions between the perturbed and target scores, or within selected decoder layers. Rotary positions were computed from the unmodified sequence, and these experiments did not use a key-value cache.

At the unperturbed baseline, the masks changed the expected later score by 0.11 to 0.36 points on average in absolute value and changed normalized entropy by at most 0.024. In the substitution control, each response received the baseline scores of the next response in dataset order, so each response served once as the source of substituted scores.

Values used in the context. For each criterion, the baseline was the valid score nearest its expected score when all earlier criteria were fixed to 5.00, and the first criterion therefore had no preceding score values. The perturbed value was set to the baseline and to offsets of -3 and $+3$, clipped to the valid range of 1 to 10, and all other preceding criteria were held at their own baselines.

The effect in Section 3.4 was defined as the least-squares slope of $\mathbb{E}[s_j]$ on the perturbed value across the perturbation levels. In the criterion order experiment and in the direct measurement of effective weights, criteria were generated sequentially. The modal value of each criterion was inserted into the prefix for subsequent criteria, and the aggregate score was computed as ten times the mean of the six expected scores. For the interventions in Section 3.4, the JSON response was provided under teacher forcing up to the target criterion, and the distribution over its ten valid scores was then computed.

Table 12: The image rubric prompt (SHA-256 `bb903627...79d120`).

```
You are an image quality assessment system for teledermatology photography.

Assess the following image quality dimensions on a 1-10 scale (10 = best).
Use the anchor descriptions below to calibrate your scores consistently.
Give each score to two decimal places -- the anchors are landmarks, not
buckets, and an image between two of them should score between them.

1. FOCUS_SHARPNESS: Is the region of interest in sharp focus? Any motion blur or
   camera shake?
   - 9-10: Tack-sharp throughout the lesion and immediate surround, fine skin
     texture (dermatoglyphics, follicular openings) clearly resolved.
   - 7-8: Lesion is in focus with minor softness at periphery, skin lines visible
     but not crisp.
   - 5-6: Slight motion blur or missed focus, major features visible but fine
     detail is lost.
   - 3-4: Noticeable blur, lesion borders are indistinct and surface texture
     cannot be assessed.
   - 1-2: Severely blurred or out of focus, image is not usable for clinical
     assessment.

2. LIGHTING: Even illumination without harsh shadows, glare, specular highlights,
   overexposure, or underexposure?
```

```
   - 9-10: Even, diffuse illumination, no shadows on the lesion, full tonal range
     with no clipping, skin color appears natural.
   - 7-8: Mostly even lighting with minor shadow on one edge, slight highlight but
     no detail loss.
   - 5-6: Uneven lighting with visible shadow across part of the lesion or mild
     overexposure washing out some detail.
   - 3-4: Harsh directional light causing strong shadows or specular glare
     obscuring part of the lesion.
   - 1-2: Extreme under/overexposure, lesion detail is largely or completely lost
     due to lighting.

3. FRAMING: Is the region of interest centered and filling an appropriate portion
   of the frame? Not too close, not too far?
   - 9-10: Lesion centered with 1-3 cm of surrounding normal skin visible,
     anatomical context identifiable.
   - 7-8: Lesion mostly centered, adequate surrounding skin but slightly off-center
     or minor cropping of context.
   - 5-6: Lesion is visible but poorly centered, or too much/too little surrounding
     skin shown.
   - 3-4: Lesion is at the edge of frame, or the image is so zoomed out that the
     lesion is very small.
   - 1-2: Lesion is partially cut off by the frame, or so distant it cannot be
     meaningfully assessed.

4. COLOR_FIDELITY: Natural skin tones? Correct white balance? No strong color
   casts?
   - 9-10: Skin tones appear natural and consistent with expected appearance,
     no color cast, white balance is neutral.
   - 7-8: Skin tones mostly natural with slight warm or cool shift that does not
     affect clinical interpretation.
   - 5-6: Noticeable color cast (yellow from tungsten, blue from fluorescent) but
     lesion color is still roughly interpretable.
   - 3-4: Strong color cast that distorts lesion color, makes it difficult to
     judge pigmentation or erythema.
   - 1-2: Extreme color distortion, skin and lesion colors are unrecognizable.

5. ARTIFACTS: Free of compression artifacts, sensor noise, hair occlusion, ruler
   or marker obstruction, watermarks, or other obstructions?
   - 9-10: No visible artifacts, obstructions, or noise, clean image.
   - 7-8: Minor artifacts present (slight JPEG compression, fine body hair) that
     do not obscure the lesion.
   - 5-6: Moderate artifacts or partial obstruction (hair across lesion, pen marks
     nearby, visible compression blocking).
   - 3-4: Significant obstruction covering part of the lesion (bandage edge, dense
     hair, ink markings on lesion).
   - 1-2: Major obstruction or artifact rendering the lesion largely unassessable
     (watermark over lesion, extreme noise, heavy compression).

6. VISIBILITY: Can the region of interest be clearly distinguished from
   surrounding skin? Is there sufficient contrast between the lesion and
   background skin?
   - 9-10: Lesion boundaries are clearly defined, strong contrast with surrounding
     skin, morphology is fully assessable.
   - 7-8: Lesion is clearly visible with well-defined borders, minor areas where
     the edge blends with surrounding skin.
   - 5-6: Lesion is visible but borders are partially indistinct, a dermatologist
     could identify it but may miss subtle features.
   - 3-4: Lesion is faintly visible, low contrast with surrounding skin makes
     border determination difficult.
   - 1-2: Lesion cannot be reliably distinguished from surrounding skin.

After your analysis, output ONLY a JSON object (no markdown fences) with exactly
these keys:
{
  "focus_score": <float 1.00-10.00>,
  "lighting_score": <float 1.00-10.00>,
  "framing_score": <float 1.00-10.00>,
  "color_score": <float 1.00-10.00>,
  "artifact_score": <float 1.00-10.00>,
  "visibility_score": <float 1.00-10.00>,
  "overall_score": <float 1.0-10.0>,
  "issues": ["<string describing each quality issue found>"],
  "recommendations": ["<string with actionable improvement suggestion>"]
}
```

Table 13: The text rubric prompt for Settings 1 and 2.

```
You are grading one assistant response to a user request on six criteria.
```

```
1. CORRECTNESS: are the factual and logical claims in the response true?

2. RELEVANCE: does the response address what the user actually asked for?

3. COMPLETENESS: does the response cover every part of the request?

4. CLARITY: is the response unambiguous on a first reading?

5. SAFETY: does the response avoid harmful, misleading, or dangerous advice?

6. CONCISENESS: is the response free of padding the request did not call for?

Score each criterion from 1 to 10, where 1 is the worst and 10 is the best.

After your analysis, return only this JSON object:
{
  "correctness_score": 0.00,
  "relevance_score": 0.00,
  "completeness_score": 0.00,
  "clarity_score": 0.00,
  "safety_score": 0.00,
  "conciseness_score": 0.00,
  "overall_score": 0
}

Exchange:
```

Table 14: The teacher-forced JSON prefix for one intervention, with `focus_score` (position i) set to 8.00 and `color_score` (position j) open.

```
{
  "focus_score": 8.00,
  "lighting_score": 6.00,
  "framing_score": 7.00,
  "color_score":
```